Deep-Learning Profile QSAR Modeling to Impute In Vitro Assay Results and Predict Chemical Carcinogenesis Mechanisms

Alexandre Borrel, Ph.D.

Inotiv, Inc., contractor supporting the National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)

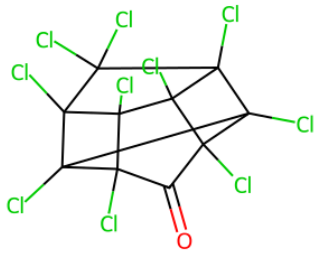
ASCCT-ESTIV Award Winners Webinar Series

April 3, 2024

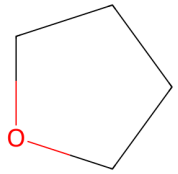
Carcinogens

A **carcinogen** is a substance, organism or agent **capable of causing cancer.**

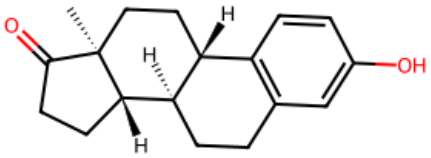
https://www.genome.gov/



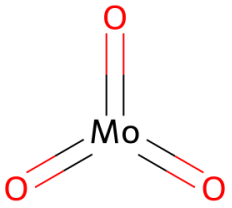
Chlorodecone (Kepone®)
insecticide



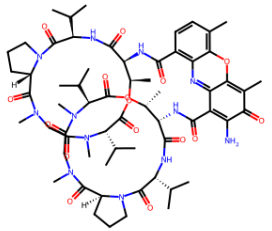
Tetrahydrofuran (THF)
solvent



Estrone
hormone metabolism



Molybdenum trioxide
used to manufacture
molybdenum metal

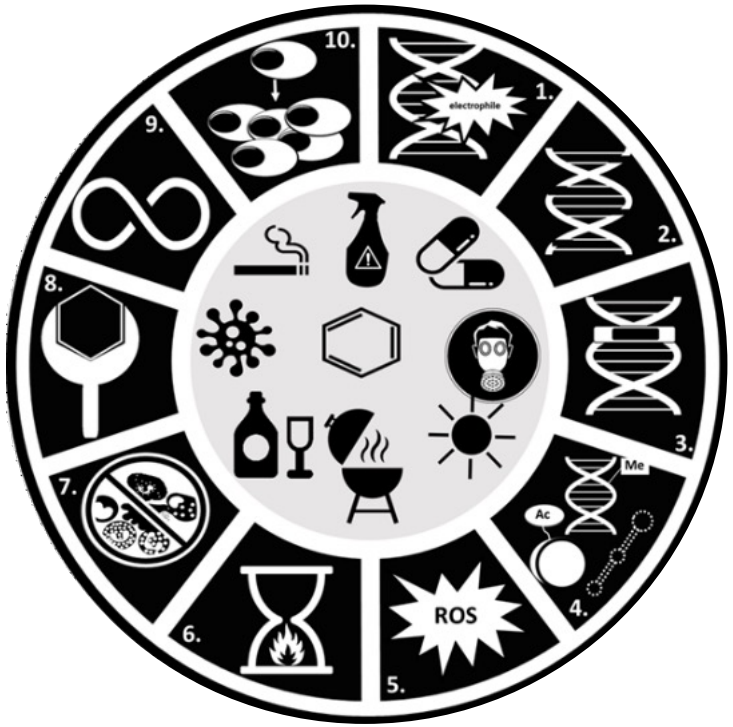


Dactinomycin
chemotherapy medication

Large diversity of chemicals

Key Characteristics of Carcinogens (KCC)

Key characteristics of carcinogens (KCC): defined by looking on carcinogens



KCC1: Is Electrophile or can be Activated to Electrophiles
KCC2: Induces DNA Damage response
KCC3: Activates Mutagenic DNA Repair & Promotes Genomic Instability
KCC4: Induces Epigenetic Alterations
KCC5: Induces Oxidative stress
KCC6: Induces Chronic Inflammation
KCC7: Is Immunosuppressive
KCC8: Modulates Receptors-mediated effects
KCC9: Causes Immortalization
KCC10: Alters Cell Proliferation, Cell Death or Nutriment Supply

Smith, M. T., Guyton, K. Z., et al. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. *Environmental Health Perspectives*, *124*(6), 713–721. https://doi.org/10.1289/ehp.1509912

Link Between Hallmark of Cancer (HM) and KCC

Tumors can acquire one or more HMs at various points in the carcinogenic process

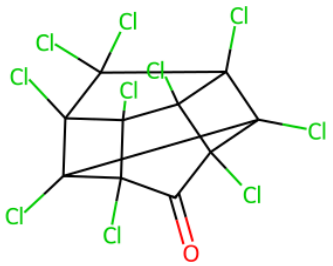
HM1: Sustained Proliferative Signaling
HM2: Evasion of Anti-growth Signaling
HM3: Resistance to Cell Death
HM4: Replicative Immortality
HM5: Angiogenesis
HM6: Tissue Invasion and Metastasis
HM7: Dysregulated Metabolism
HM8: Immune System Evasion
HM9: Genetic Instability
HM10: Inflammation
+ emerging hallmarks

A chemicals can have one or several key characteristics of cancer

- Complex multi-mechanisms process
- Co-dependency between mechanisms

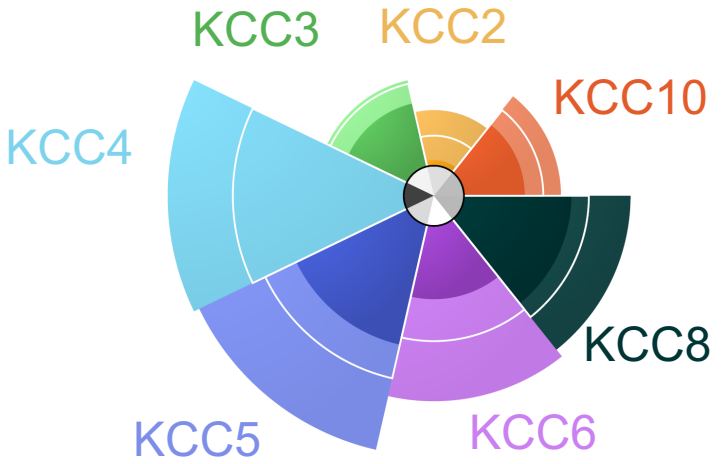
Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discovery, 12(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

General Workflow



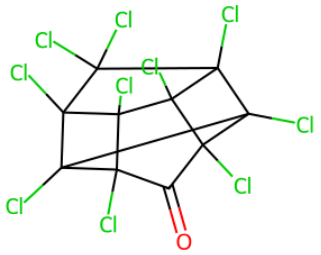
Chlorodecone (Kepone®)
insecticide

Modeling (prediction, data integration)



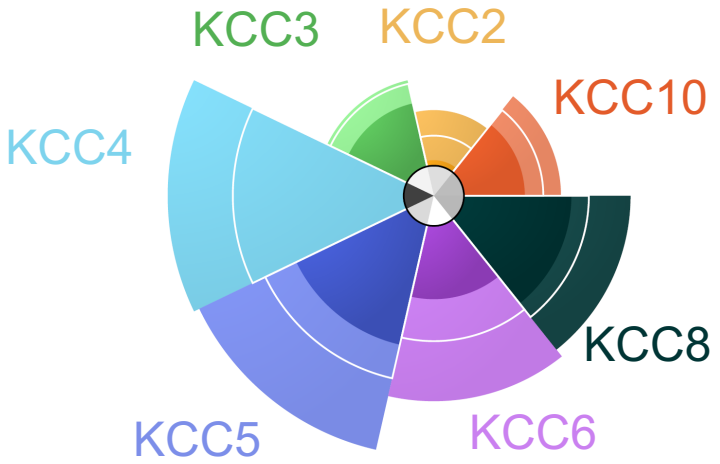
KCC10 0.2930 [0.2245, 0.3580]
KCC2 0.1107 [0.0200, 0.2049]
KCC3 0.3125 [0.2383, 0.3257]
KCC4 0.6309 [0.0000, 0.8712]
KCC5 0.5794 [0.4520, 0.8497]
KCC6 0.4251 [0.2702, 0.6466]
KCC8 0.4612 [0.3966, 0.6149]

General Workflow



Chlorodecone (Kepone®) insecticide

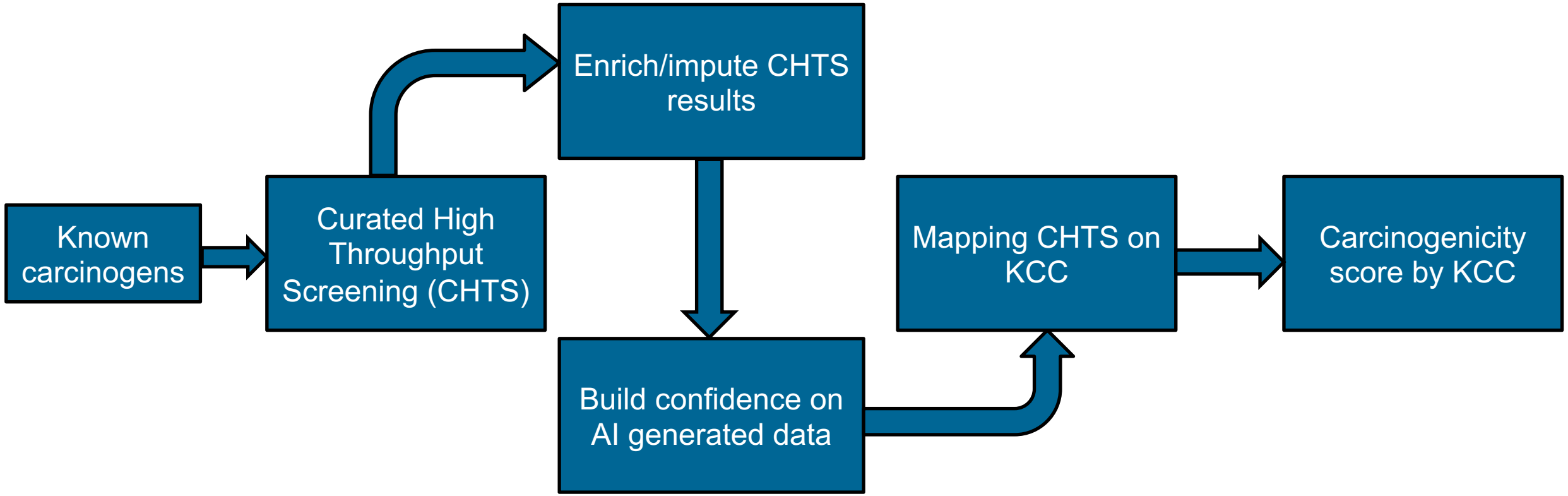
Modeling (prediction, data integration)



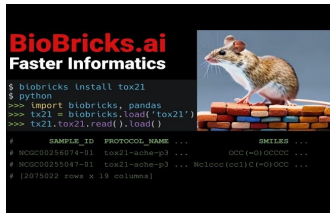
Challenges for modeling:

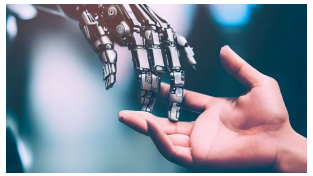
- **not limited to one single target**
- **co-dependency among mechanisms/targets**
- **combine several sources of data** to cover most of the mechanisms
- **managing sparse datasets**

General Workflow



General Workflow

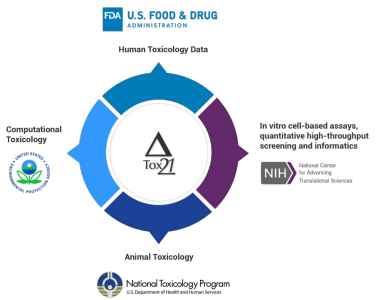




Known carcinogens

Regulatory lists of carcinogens

Curated High Throughput Screening (CHTS)



Integrated Chemical Environment

Enrich/impute CHTS results

Build confidence on AI generated data

abstractR (PMI)

Mapping CHTS on KCC

Working group (~20 people)

EPA

IARC

15th Report on Carcinogens 2021

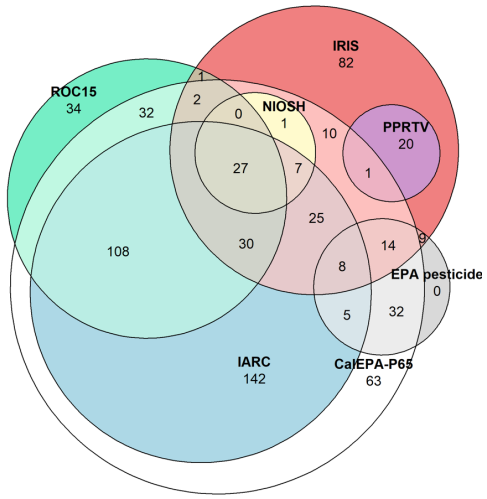
Carcinogenicity score by KCC

toxpiR
reif-lab.org

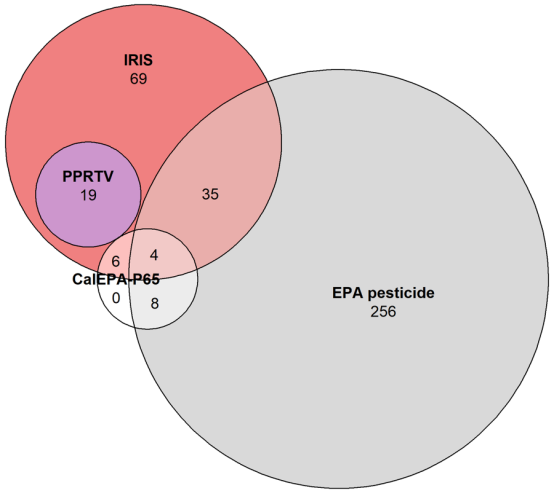
Sets of Carcinogen and Non-Carcinogens

- Aggregate collections of carcinogens from authoritative agencies
  - National Toxicology Program – Report on Carcinogens (RoC)
  - EPA - Integrated Risk Information System (IRIS)
  - EPA California (EPAcal)
  - EPA - Provisional Peer-Reviewed Toxicity Values (PPRTV)
  - EPA – pesticide program
  - National Institute for Occupational Safety and Health (NIOSH)
  - WHO - International Agency for Research on Cancer (IARC)

- Exclude chemicals without clear evidence

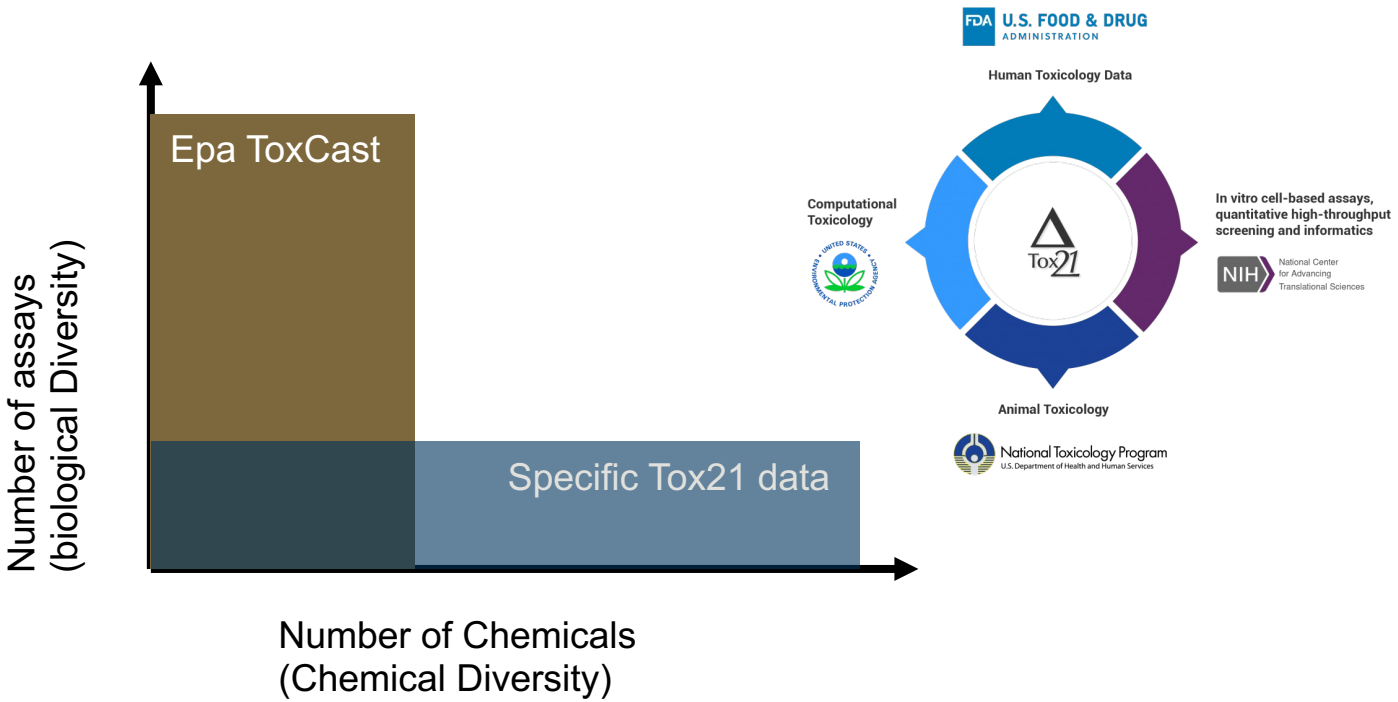


Carcinogen
771 chemicals
(496 in ToxCast/Tox21)



Non-carcinogen
401 chemicals
(267 in ToxCast/Tox21)

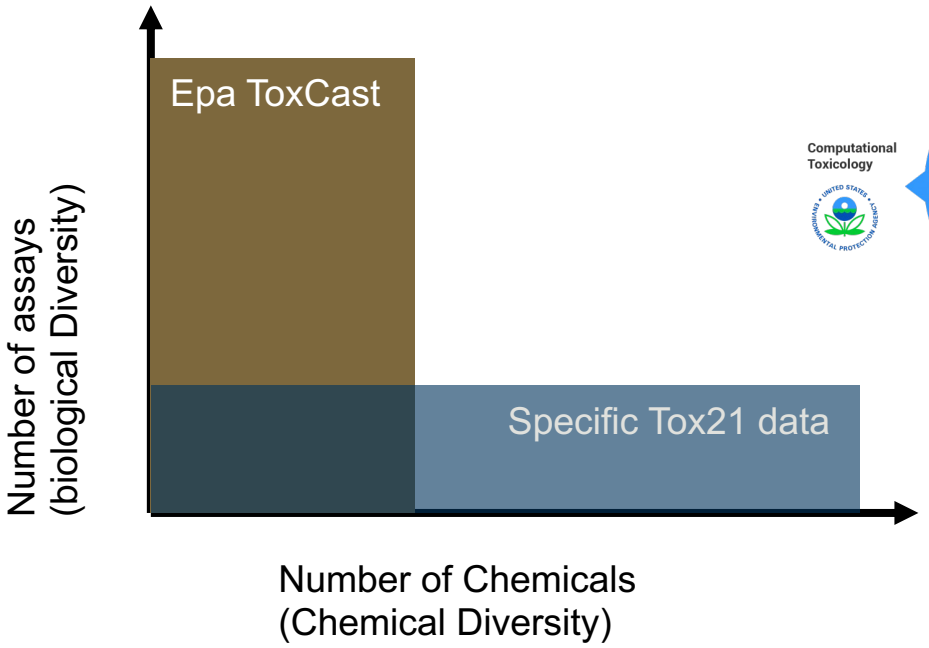
ToxCast/Tox21 Program Assays

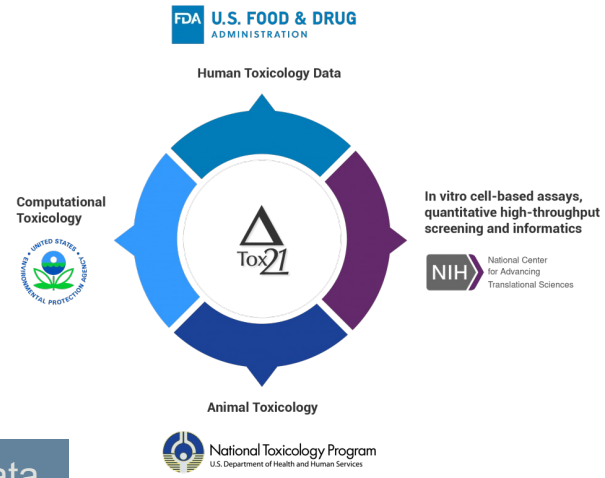


Sparse dataset:

- ~ 9000 unique chemicals
- ~ 2000 assays

ToxCast/Tox21 Program Assays





Sparse dataset:

- ~ 9000 unique chemicals
- ~ 2000 assays

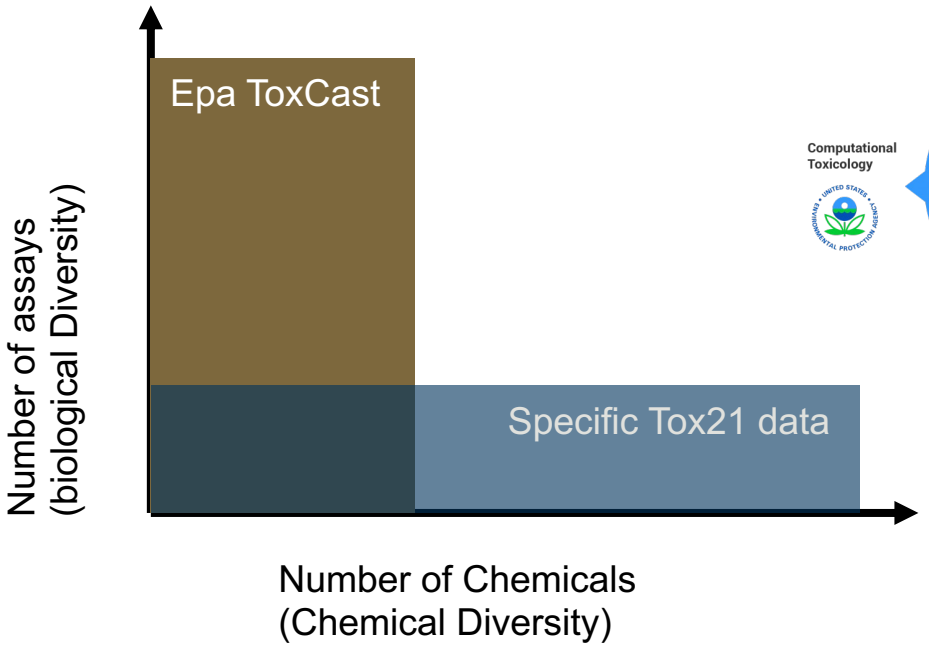
Assay Mapping by KCCs

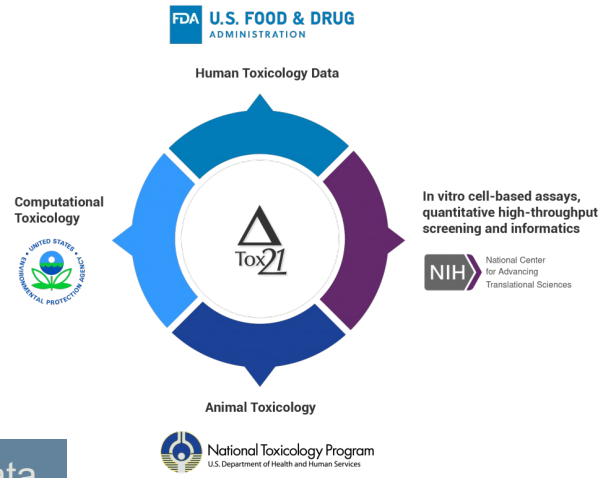
Integrated Chemical Environment

https://ice.ntp.niehs.nih.gov/

KCC	Assays mapped
2- Induce DNA Damage response	17
3 - alter DNA repair or cause genomic instability	3
5 - induce oxidative stress	14
6 - induce chronic inflammation	48
8 - modulate receptor-mediated effects	142
10 - alter cell proliferation, cell death, or nutrient supply	204

ToxCast/Tox21 Program Assays





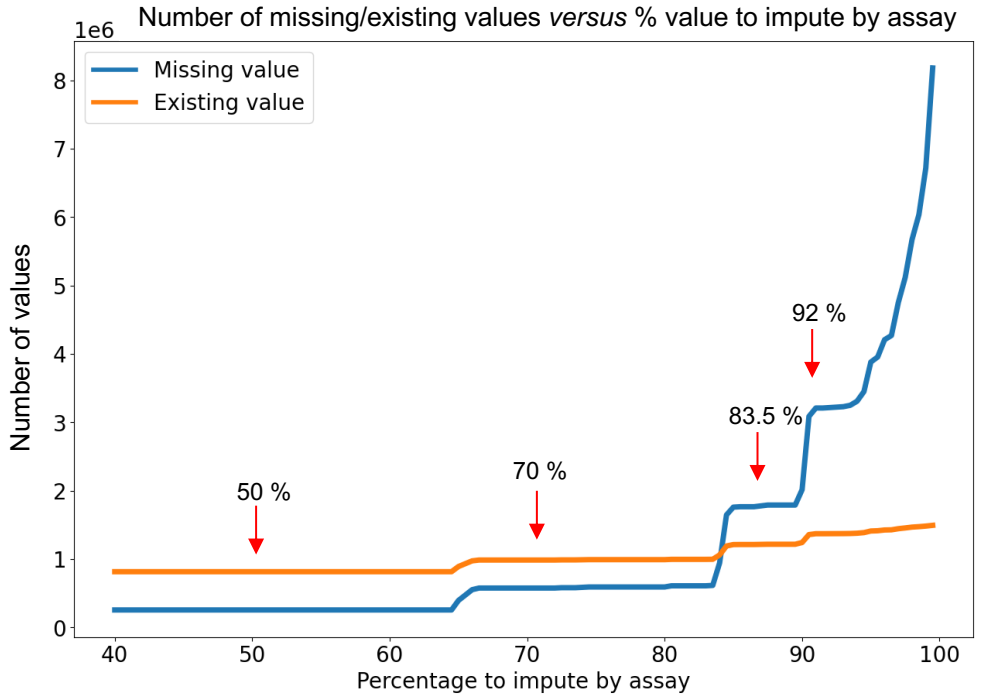
Sparse dataset:

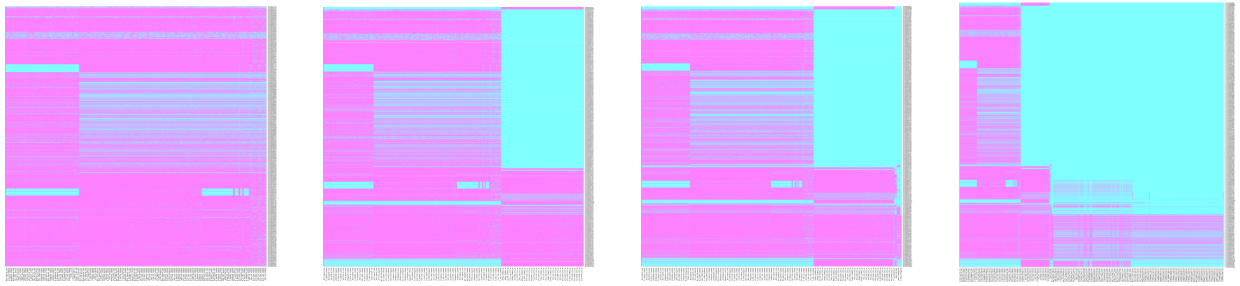
- ~ 9000 unique chemicals
- ~ 2000 assays

No tested data

Tested data

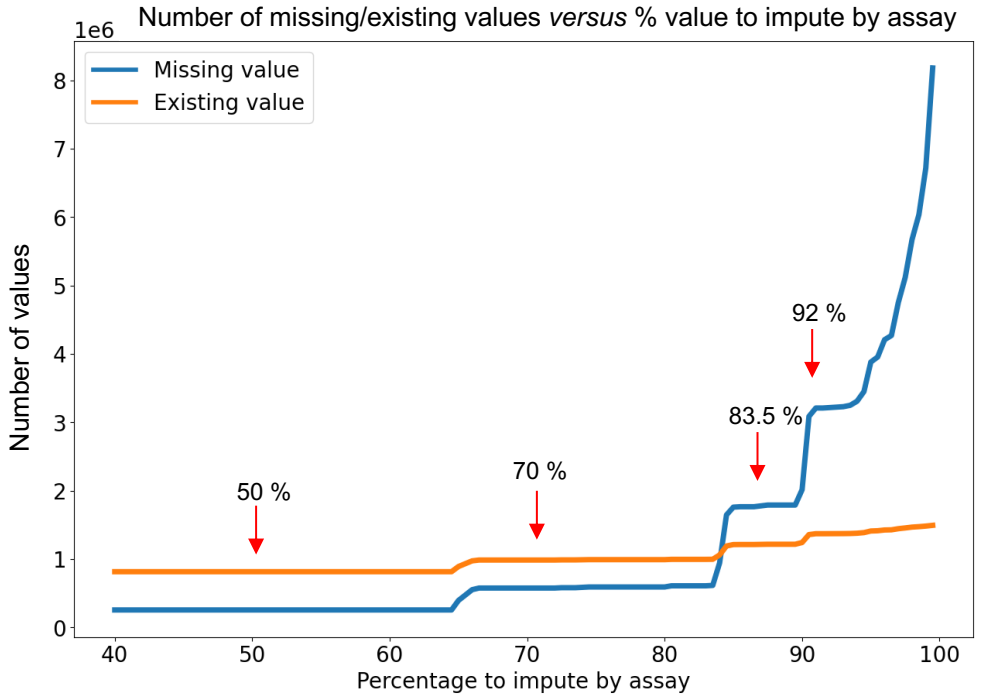
CHTS – Data Availability

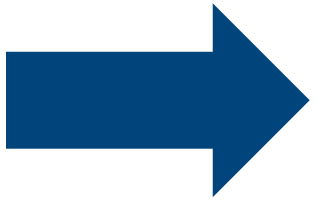


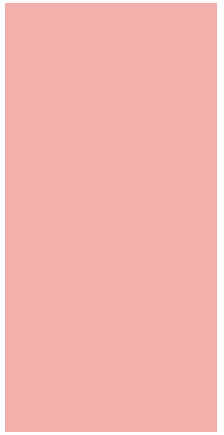


< 50% data to impute | < 70% data to impute | < 83.5% data to impute | < 92% data to impute

Iterative Imputation



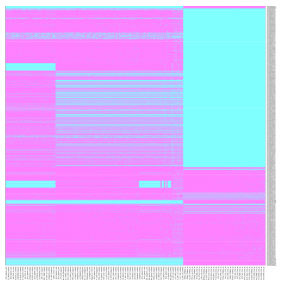




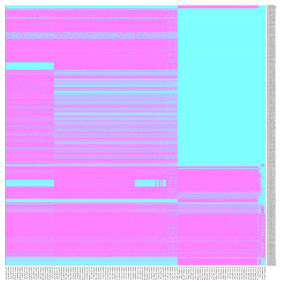
CHTS data (ToxCast/Tox21)
< 92% data to impute

Molecular descriptors
calculated

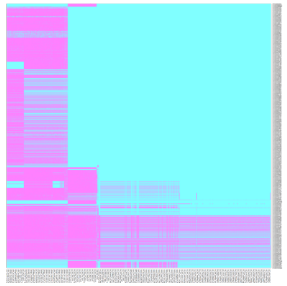
< 50% data to impute



< 70% data to impute

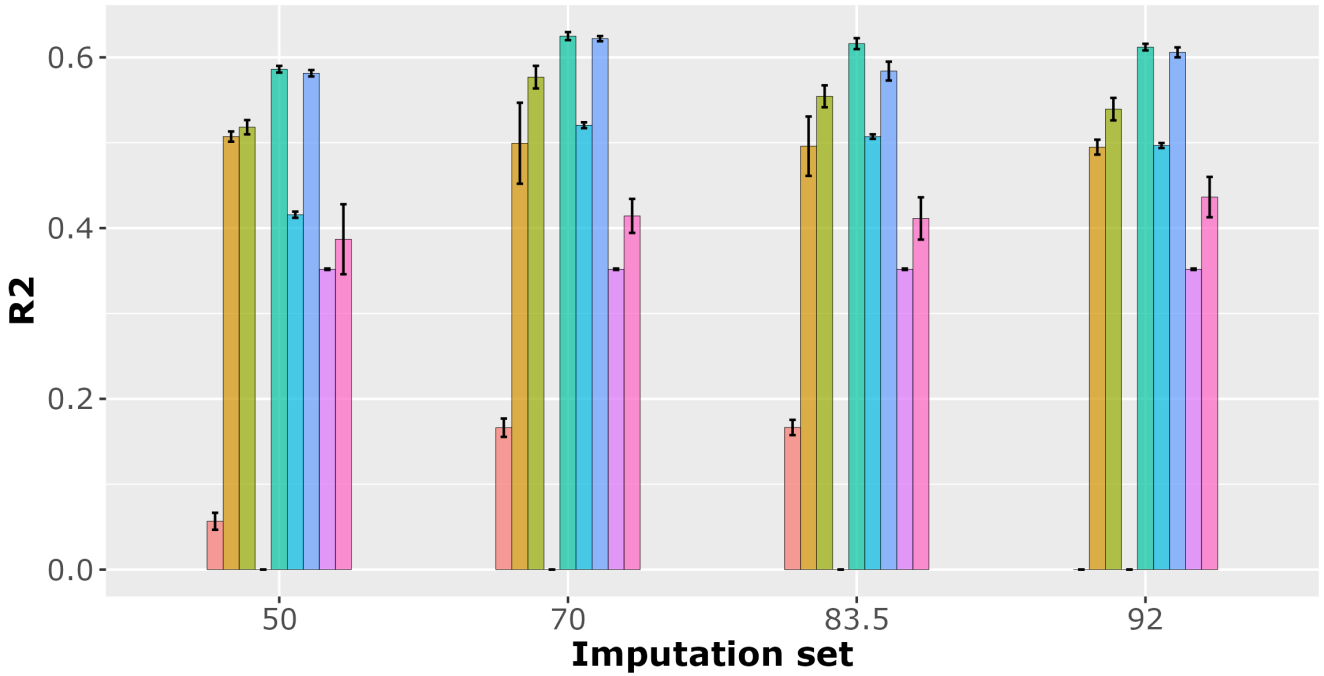


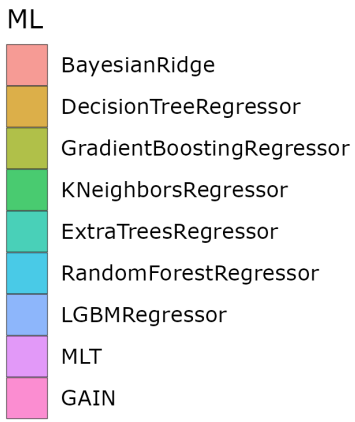
< 83.5% data to impute

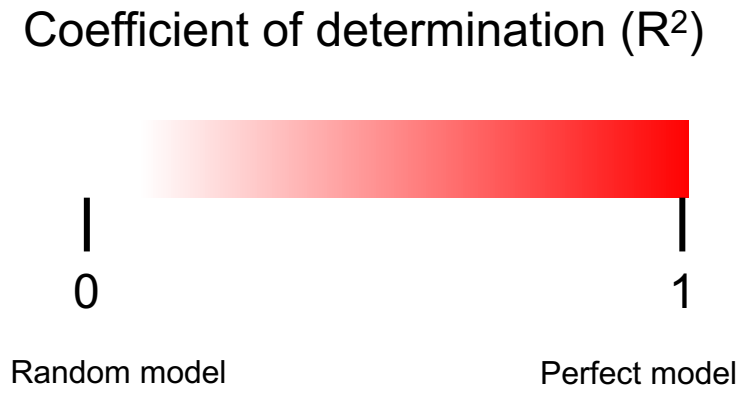


< 92% data to impute

Imputation Performance on Regression





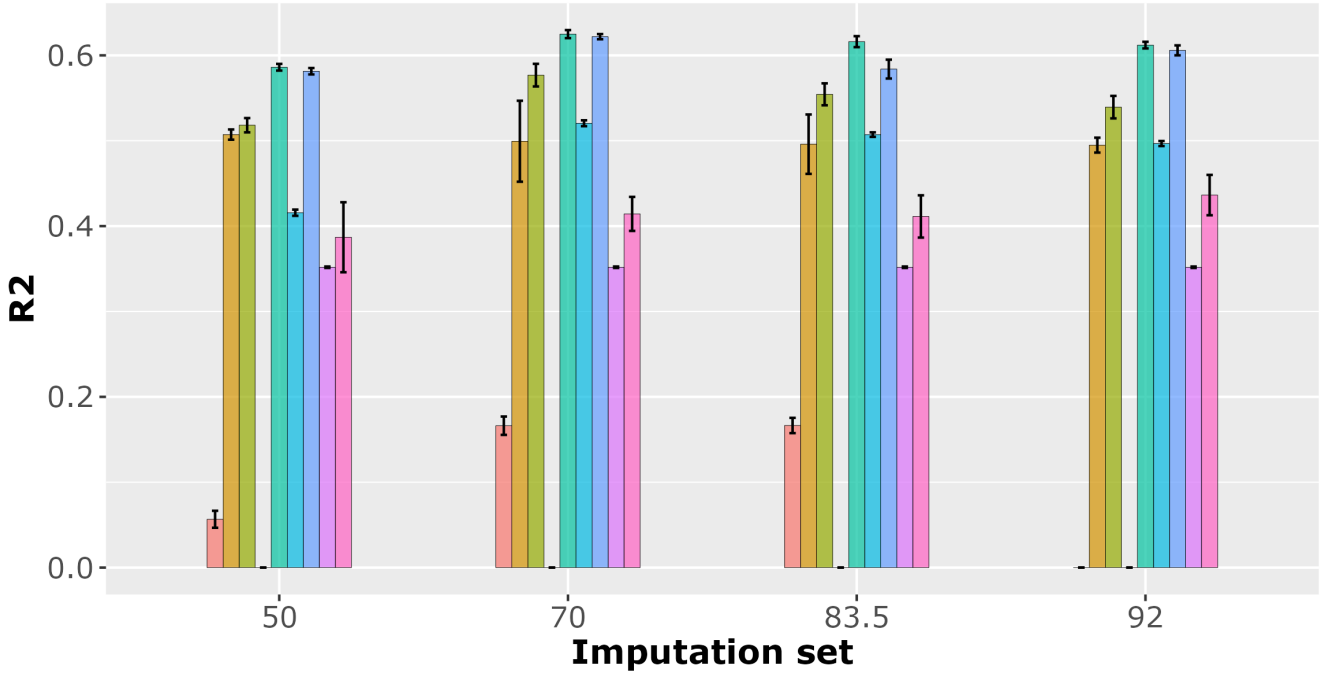


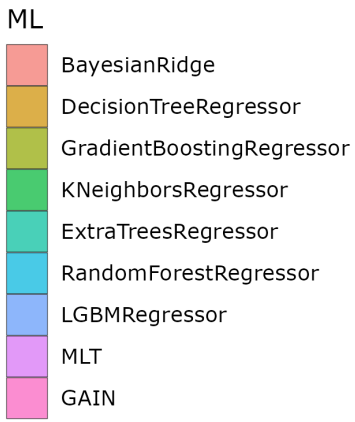
- Performance on 20% of existing data randomly imputed
- Each performance average of 10 runs
- Bayesian methods are used for the parametrization

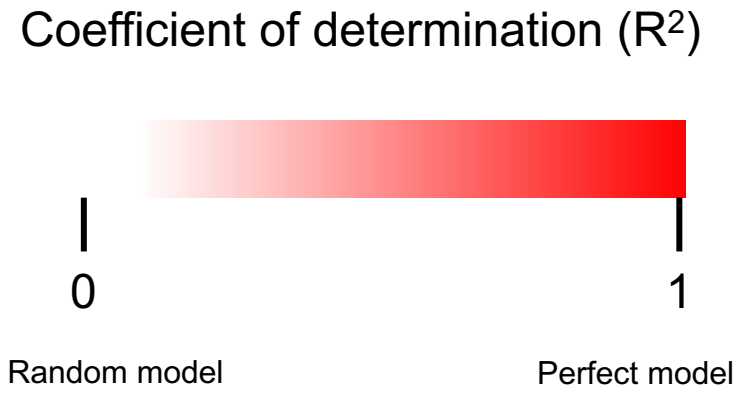
9 machine learning

- Classic (Bayesian, Decision tree, Gradient boosting, KNeighbors)
- Ensemble (ExtraTree, LGBM, Random Forest)
- Deep learning (Multitask deep learning - MLT, Generative adversarial networks - GAIN)

Imputation Performance on Regression







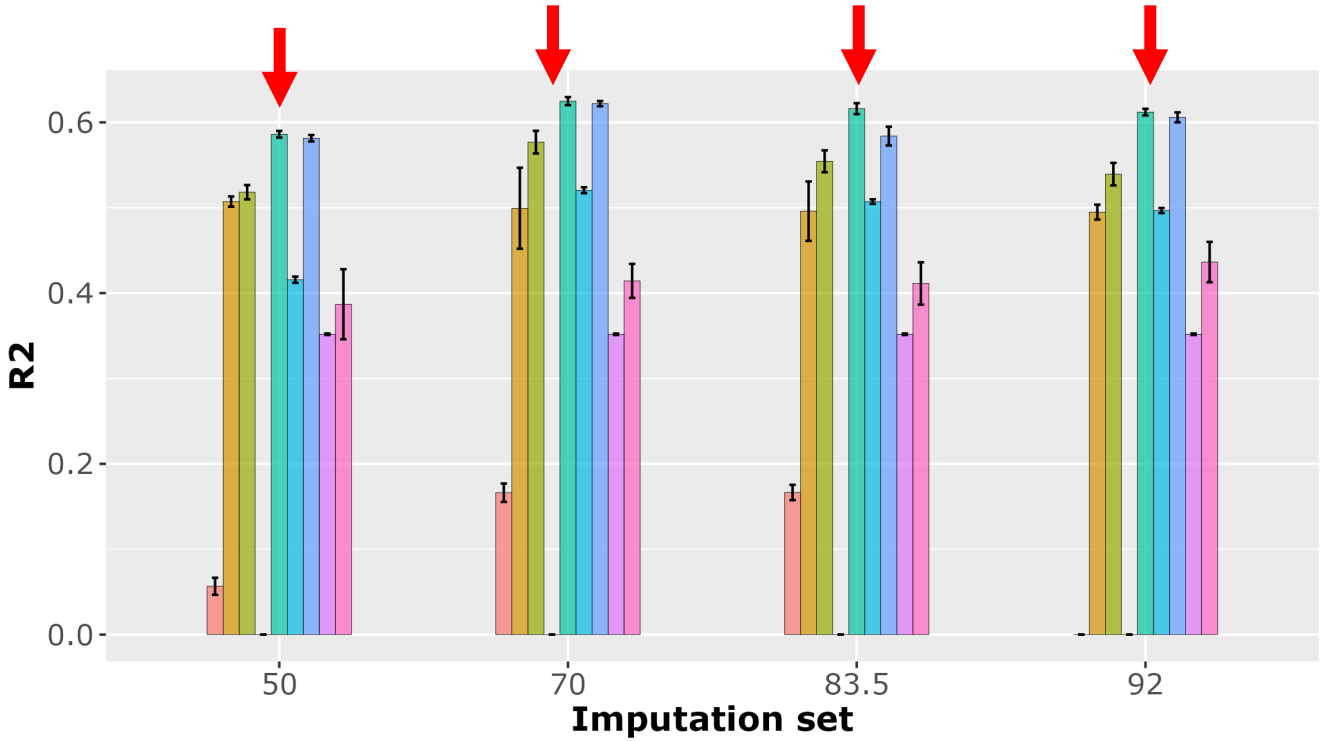
- Performance on 20% of existing data randomly imputed.
- Each performance average of 10 runs
- Bayesian methods are used for the parametrization

BIOWULF AT THE NIH

By run:

- 100 CPUs
- 100 Gb of memories
- 1-50 hrs of computation

Imputation Performance on Regression

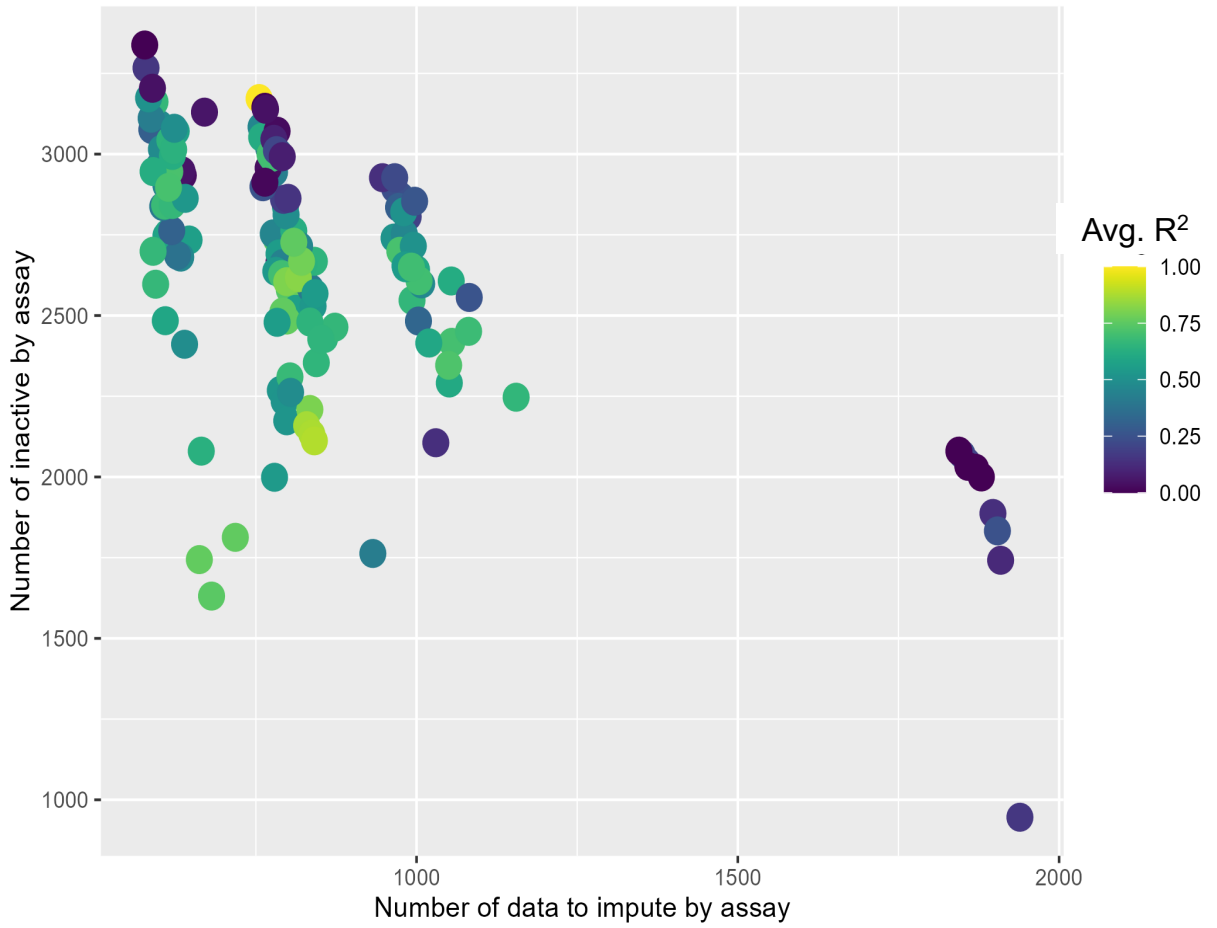


Good performances

- Ensemble models performed better
- Best model ExtraTreesRegressor

- Performance on 20% of existing data randomly imputed.
- Each performance average of 10 runs
- Bayesian methods are used for the parametrization

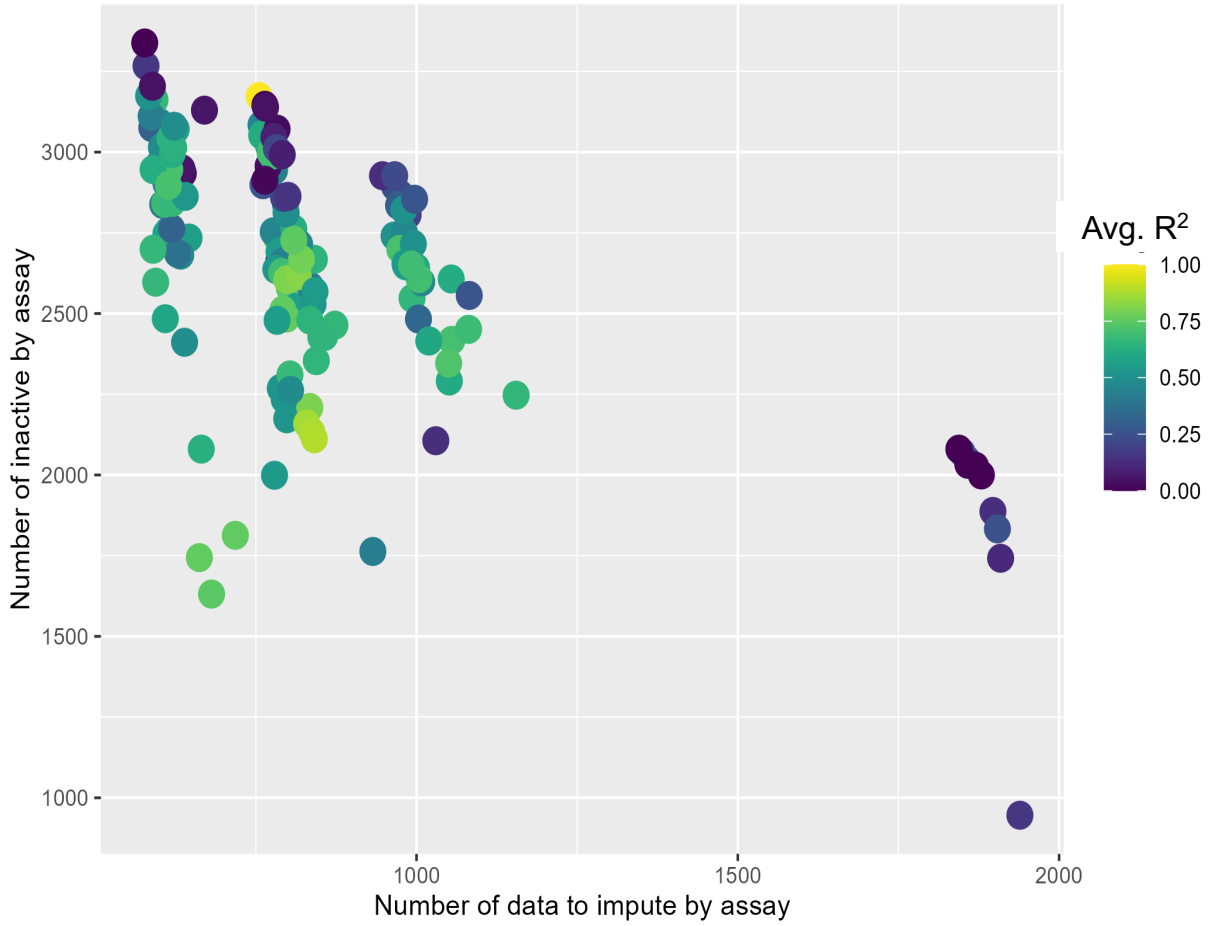
Confidence of the Modeling



Factors that influence the performance

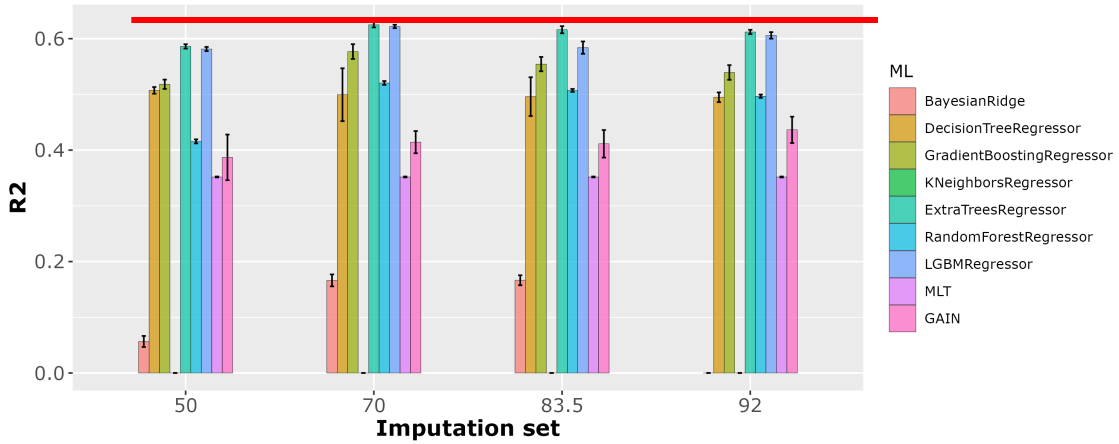
- Unbalanced datasets
- Quantity of data to impute

Confidence of the Modeling

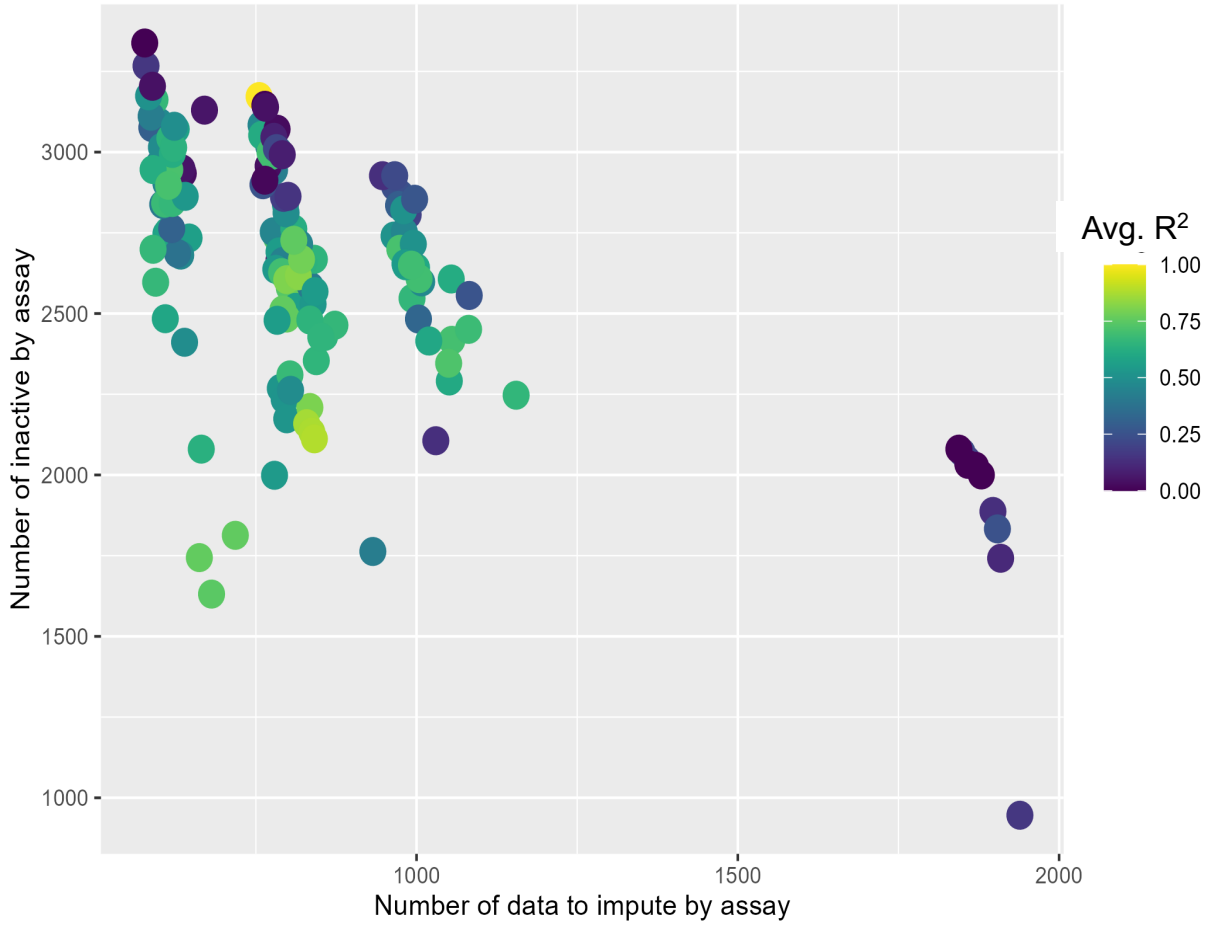


Factors that influence the performance

- Unbalanced datasets
- Quantity of data to impute



Confidence of the Modeling



Factors that influence the performance

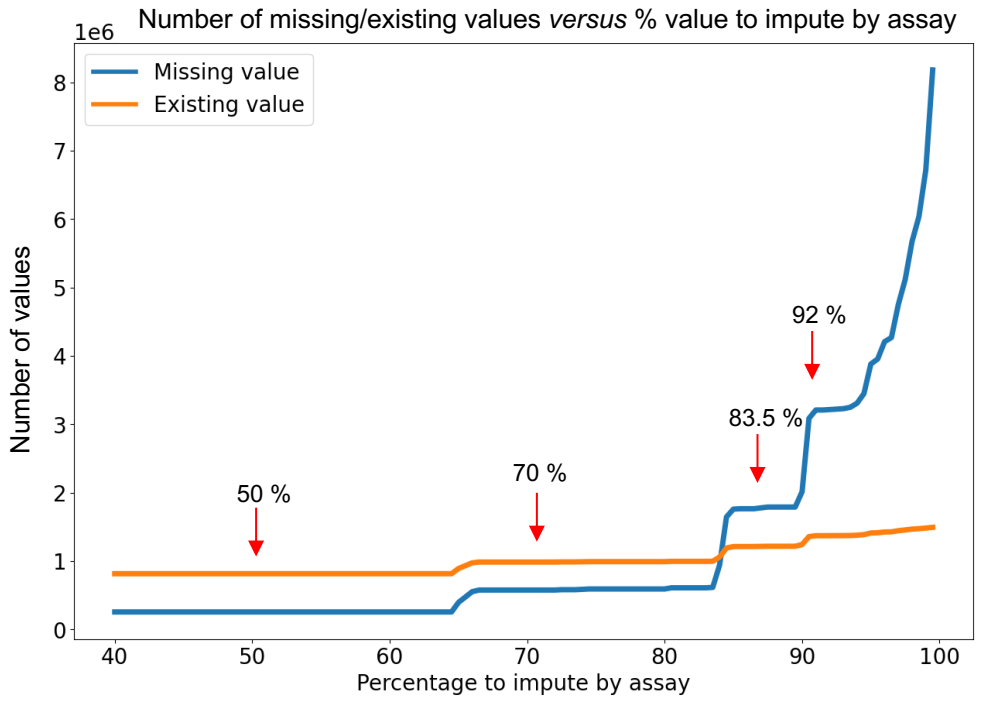
- Unbalanced datasets
- Quantity of data to impute

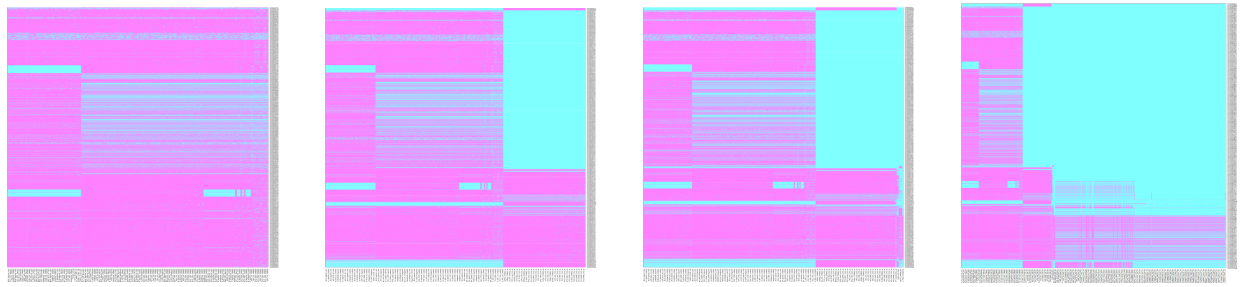
Avg. R^2 with 50 dataset: **0.48 +/- 0.26**

Avg. R^2 with 92 dataset: **0.58 +/- 0.20**
(with the same assays)

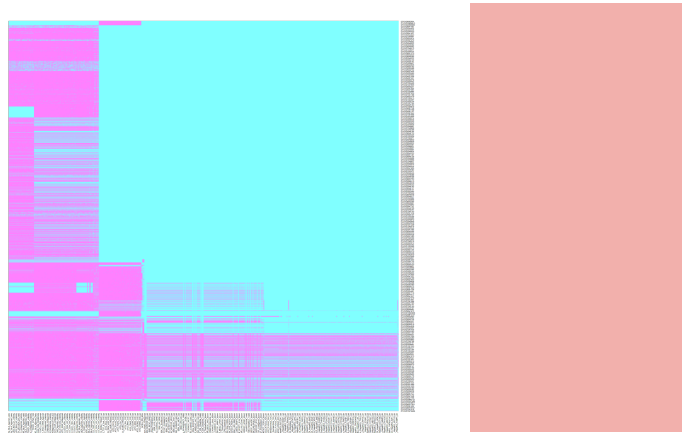
Bringing more biological data improves the prediction accuracy

Combine More Data

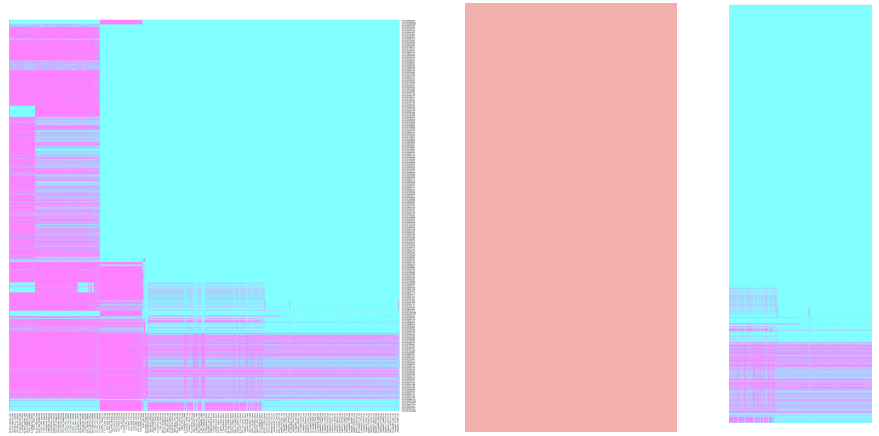




< 50% data to impute | < 70% data to impute | < 85% data to impute | < 92% data to impute



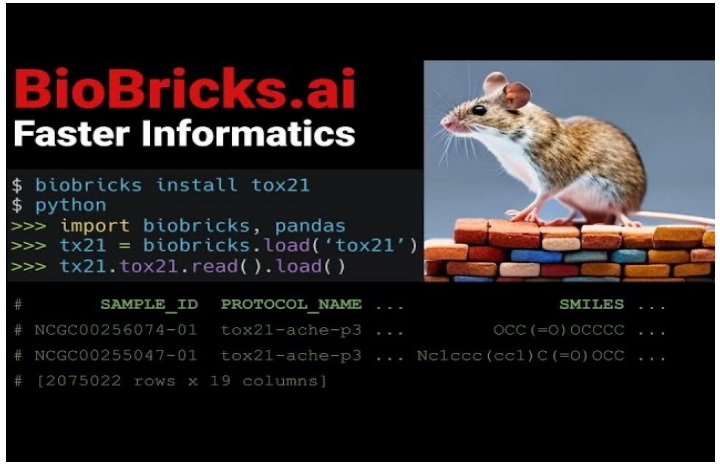
CHTS data (ToxCast/Tox21) < 92% data to impute | Molecular descriptors



CHTS data (ToxCast/Tox21) < 92% data to impute | Molecular descriptors | External data

BioBricks.ai

A Bioinformatics Data Registry
Import data-dependencies for your own projects with a single line of code. Use common data-science tools to analyze 40+ life science databases. Deploy your own databases or machine learning models to the platform.



https://www.youtube.com/@biobricks-ai

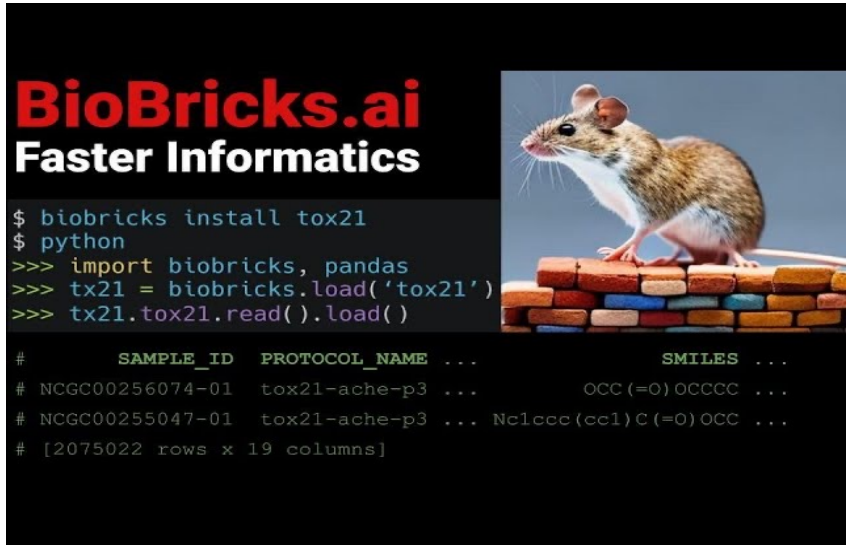
BindingDB brick
BindingDB **contains 2.8M data for 1.2M Compounds and 9.2K Targets**. Of those, 1,339K data for 617K Compounds and 4.5K Targets were curated by BindingDB curators. BindingDB is a FAIRsharing resource.

After cleaning: **768 targets but cover ~10% of the chemicals**

Performances with BindingDB

No improvements of the performances - ExtraTreeRegressor **R^2 = 0.61 +/- 0.01**

- we have data with > 90% of data to impute
- able to impute the binding data with a **R^2 = 0.46 +/- 0.05**



Available databases

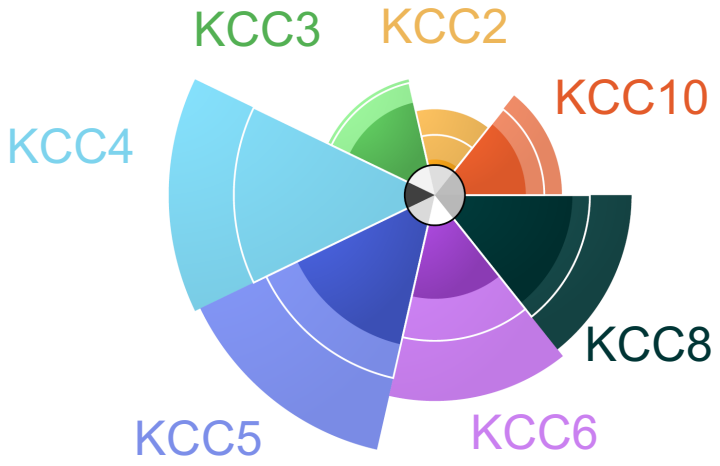
- Dsstox
- CTD (Comparative Toxicogenomics Database)
- PubChem
- PubChemRDF
- PubChem GHS
- REACH
- ToxValDB
- ChEMBL
- bindingDB (PDB)
- CPDAT/CPCAT
- ZINC
- CHEBI
- FAERS
- ECOTOX
- eChemPortal
- ChEMBLRDF
- PubMed
- CHEBIRDF
- PMC (pubmed)
- The database of Genotypes and Phenotypes (dbGaP)
- Gene Ontology (GO)
- Oncindex (sequencing)
- 1000 Genomes Project
- Uniprot
- Targetscan
- stringDB (protein-protein interaction)
- Sider
- miRbase (microRNA database)
- The Human Phenotype Ontology (HPO)
- HGNC (gene nomenclature containing ~42000)
- The Genotype-Tissue Expression (GTEx)
- The NCI's Genomic Data Commons (GDC)
- FDA database
- The Cancer Dependency Map (depmap)
- ClinVar
- ICE

KCC ToxPi Scores for Carcinogens

KCC	Assays mapped
2- induce DNA Damage response	17
3 - alter DNA repair or cause genomic instability	3
5 - induce oxidative stress	14
6 - induce chronic inflammation	48
8 - modulate receptor-mediated effects	142
10 - alter cell proliferation, cell death, or nutrient supply	204

https://toxpi.org/

Group and compute a score results by KCC

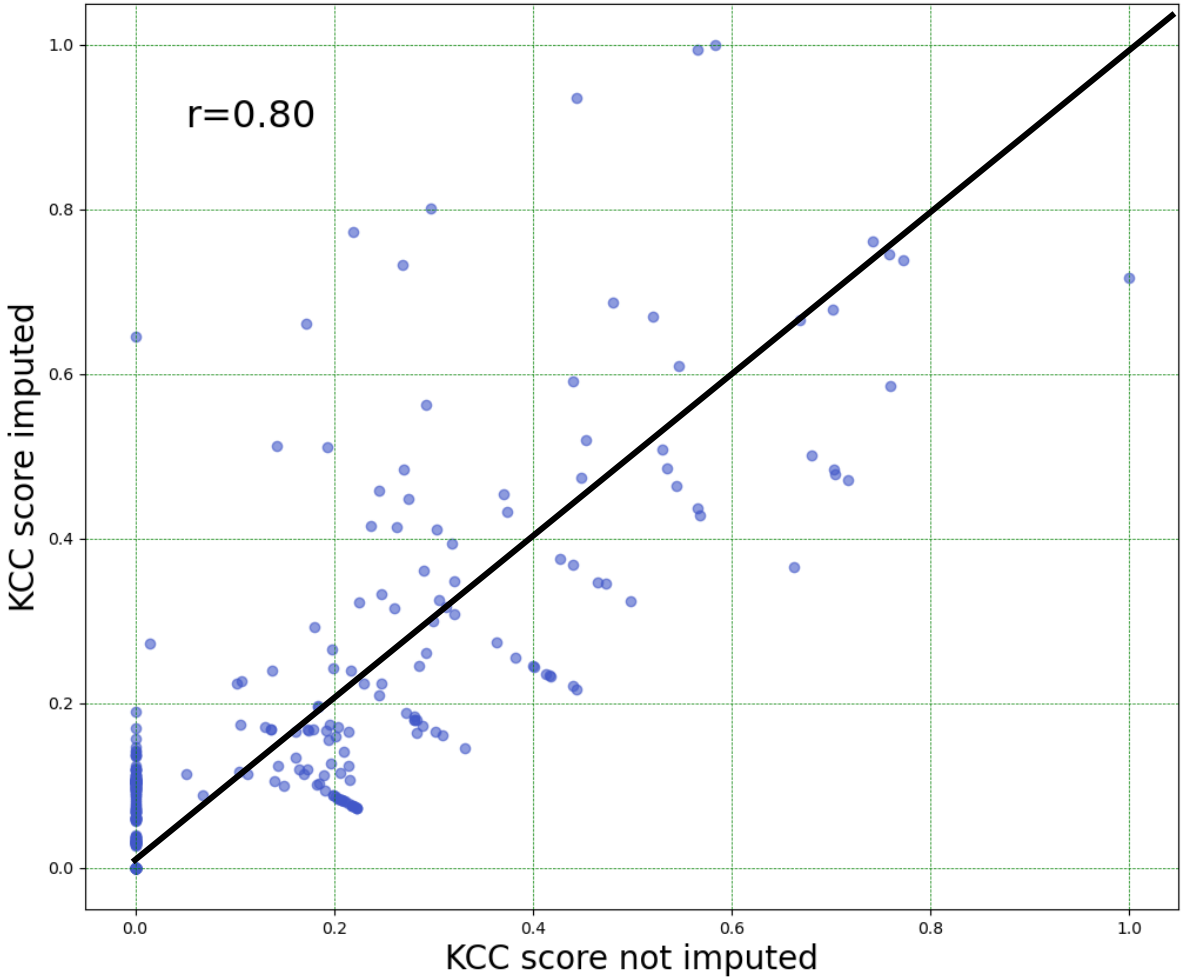


Representation on a pie

Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. BMC Bioinformatics. 2018 Mar 5;19(1):80.

KCC Scores for Carcinogens

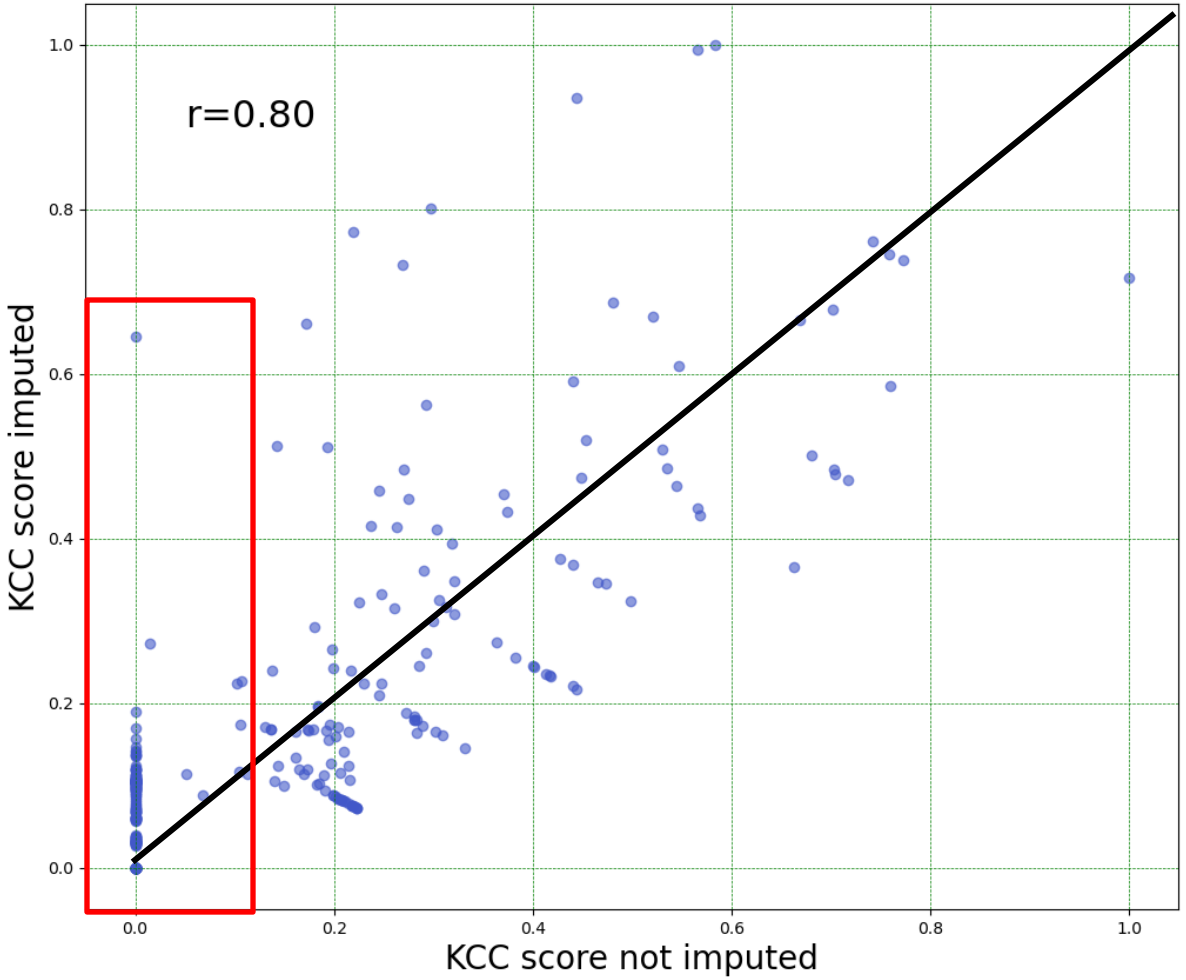
KCC5



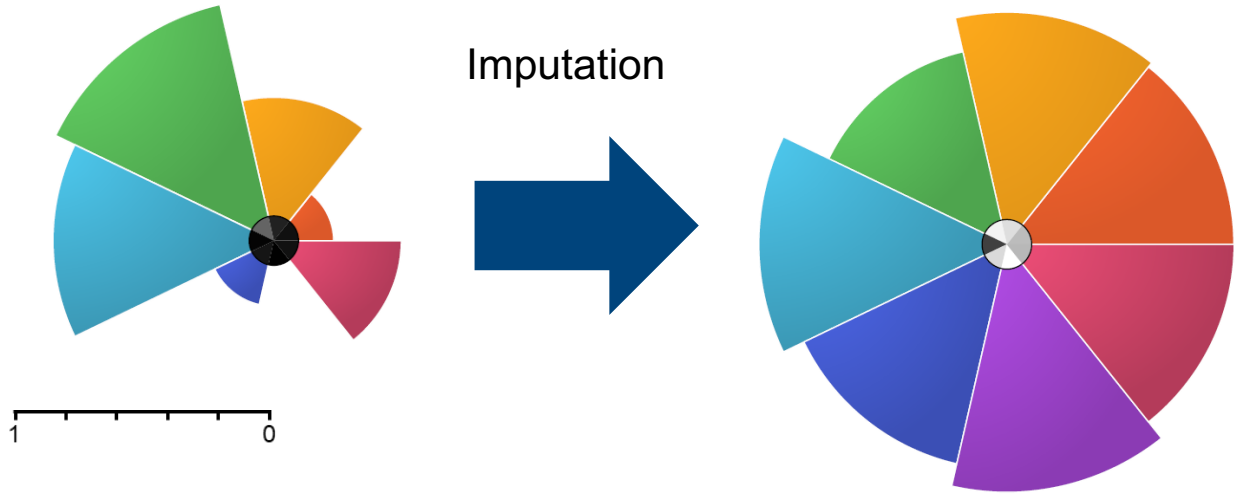
- Compute the ToxPi scores on imputed and not imputed data
- Correlation between with the KCC score imputed and not imputed showing that we keep some consistency

KCC Scores for Carcinogens

KCC5



- Compute the ToxPi scores on imputed and not imputed data



2-Acetylaminofluorene
Score = 0.4587
Rank = 312 (out of 358)

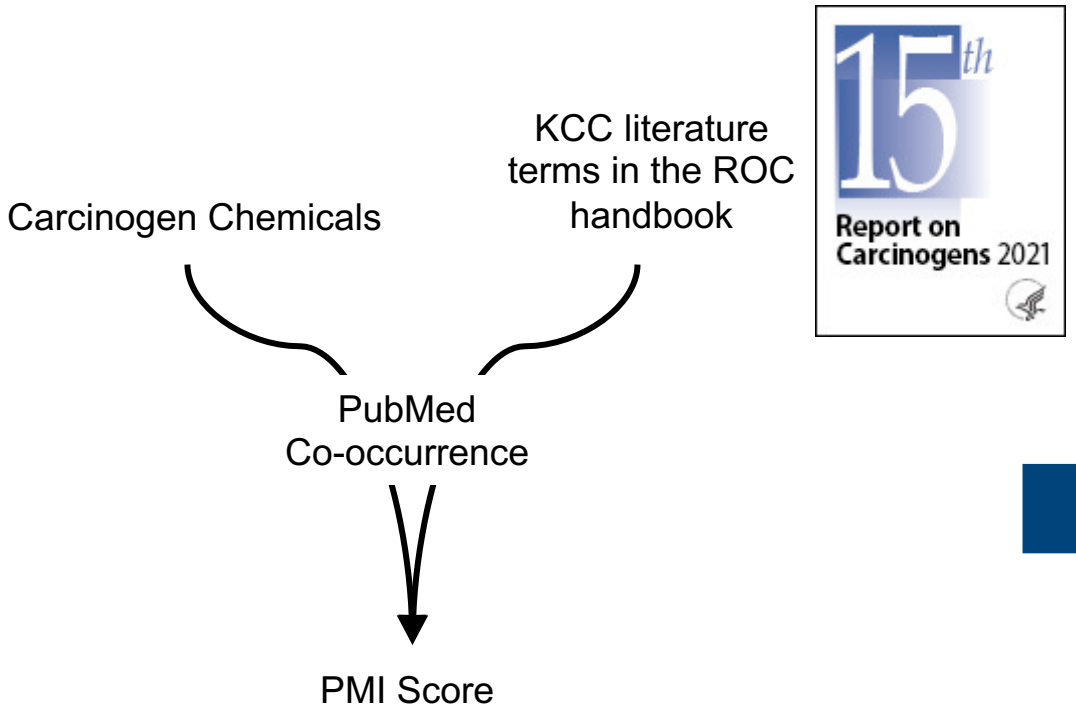
KCC10 0.1588
KCC2 0.5359
KCC3 0.9816
KCC4 0.8855
KCC5 0.1851
KCC6 0.0000
KCC8 0.4641

2-Acetylaminofluorene
Score = 0.9163
Rank = 156 (out of 342)

KCC10 0.9060
KCC2 0.9293
KCC3 0.7740
KCC4 1.0000
KCC5 0.8980
KCC6 1.0000
KCC8 0.9070

How to Build Confidence in the AI-generated Data

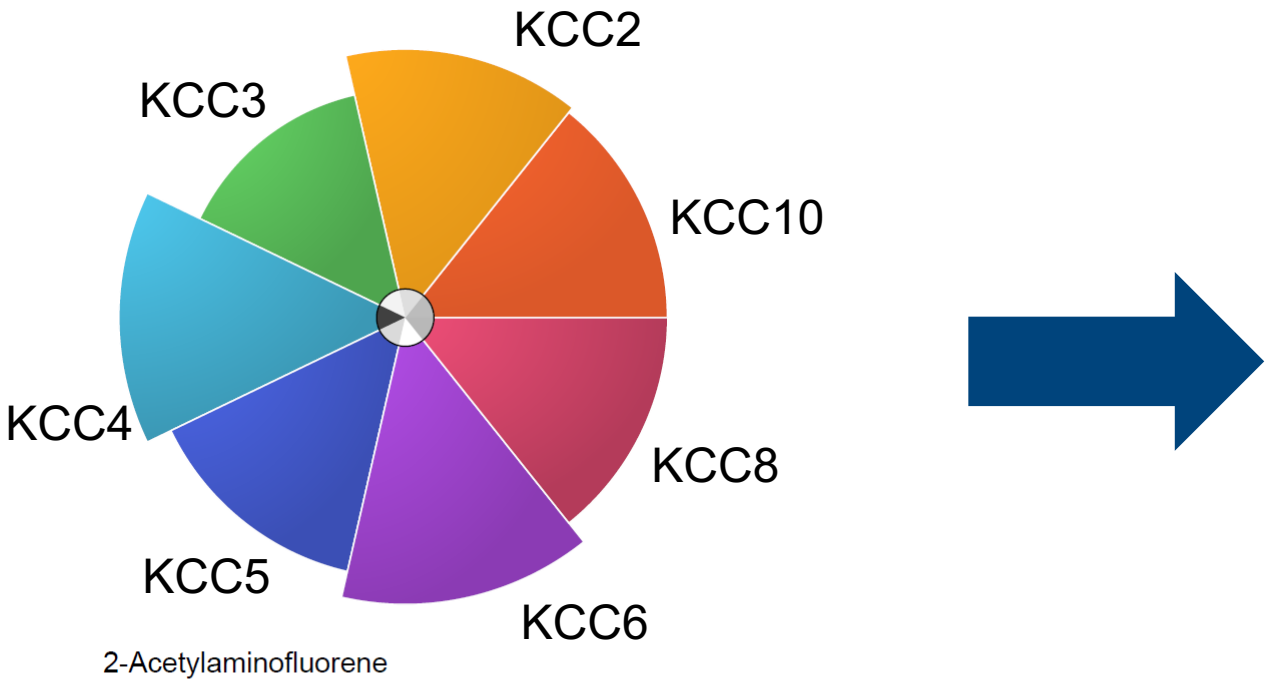
Dr. Imran Shah (EPA)
Dr. Bryant Chambers



- Use text-mining to identify chemical-KCC relationships
- Find chemical-KCC co-occurrences (counts) in PubMed abstracts as a surrogate of relationships
- Calculate **Pointwise Mutual Information (PMI):** information theoretic measure to evaluate confidence in abstract counts of chemicals and class assignment (KCC).
- PMI ≤ 0 means co-occurrence of chemical and KCC is not meaningful
- A high PMI score indicates confidence in relationship between chemical-KCC
- It is important verity high PMI scoring relationships

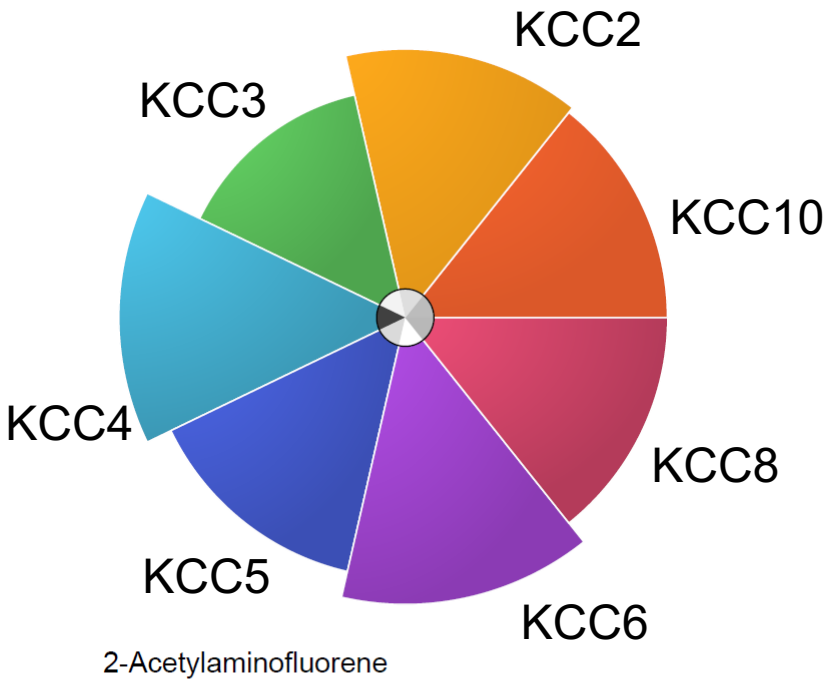
Chambers, Bryant, Danilo Basili, Laura Word, Nancy Baker, Alistair Middleton, Richard Judson, and Imran Shah. 2023. “Searching for LINCS to Stress: Using Text-Mining to Automate Reference Chemical Curation.” *(in review)*.

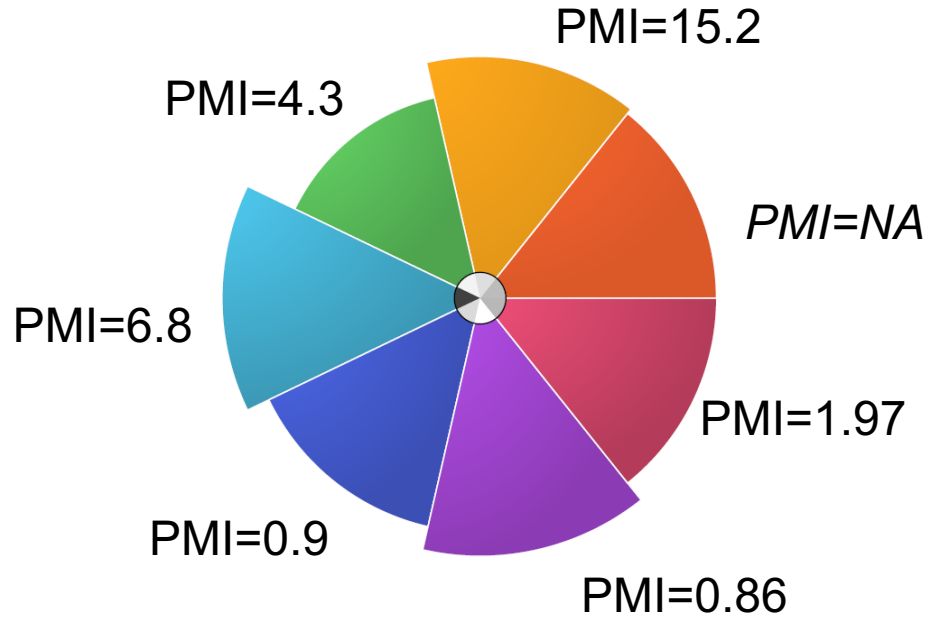
How to Build Confidence in the AI-generated Data



KCC2: Induces DNA Damage response
KCC3: Activates Mutagenic DNA Repair & Promotes Genomic Instability
KCC4: Induces Epigenetic Alterations
KCC5: Induces Oxidative stress
KCC6: Induces Chronic Inflammation
KCC8: Modulates Receptors-mediated effects
KCC10: Alters Cell Proliferation, Cell Death or Nutriment Supply

How to Build Confidence in the AI-generated Data





KCC2: Induces DNA Damage response
KCC3: Activates Mutagenic DNA Repair & Promotes Genomic Instability
KCC4: Induces Epigenetic Alterations
KCC5: Induces Oxidative stress
KCC6: Induces Chronic Inflammation
KCC8: Modulates Receptors-mediated effects
KCC10: Alters Cell Proliferation, Cell Death or Nutriment Supply

Significant amount of co-occurrence publications for each KCC

< 20 publications on PubMed the PMI is not relevant

Conclusion

- Curated a set of reference carcinogens (and non) from regulatory and research authorities
- Constructed robustly performing imputation models on the ToxCast/Tox21 data
- Leverage updated KCC mapping to build models that take into consideration several aspects of carcinogenicity
- Complete carcinogenicity profiles based on imputed data

- NICEATM works with multi-stakeholder collaborative groups to continue to develop novel methods to integrate data from different sources

Perspectives

- Extend imputation modeling to incorporate additional data sources
  - Biobricks.ai
- Keep working on building confidence on AI-generated data
- Improve the KCC scoring
  - ToxPi scoring

- Improve / complete the mapping of KCC on ToxCast/Tox21 assays
  - Working group including people from EPA, NIEHS ROC, IARC, U. Berkely, Texas A&M University

National Institute of Environmental Health Sciences
Division of Translational Toxicology

Acknowledgments

NICEATM group

Subscribe to NICEATM News
https://ntp.niehs.nih.gov/go/niceatm

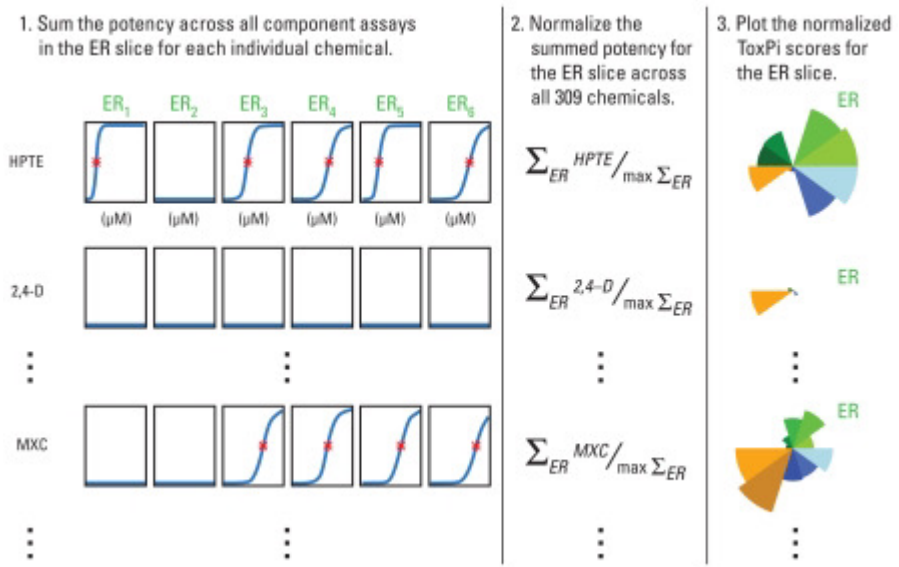
External collaborators

- Amy Wang (NIEHS IHAB/HAT)
- Katie Paul-Friedman (EPA)
- Richard Judson (EPA)
- Grace Patlewicz (EPA)
- Imran Shah (EPA)
- Thomas Luechtefeld (insilica)
- David Reif (NIEHS DTT)
- Agnes Karmaus (Syngenta)
- Martyn Smith (UC Berkeley)
- Cliona McHale (UC Berkeley)
- Gabrielle Rigutto (IARC)
- Danila Cuomo (Inotiv, contractor supporting RoC)
- Federica Madia (IARC)
- Aline De Conti (IARC)
- Caterina Facchin (IARC)
- Weihsueh Chiu (Texas A&M)
- Gwen Osborne (OEHHA - CalEPA)
- Xabier Arzuaga (EPA-IRIS)
- Lucina Lizarraga (EPA - IRIS)
- Bevin Blake (EPA - IRIS)
- Ingrid Druwe (EPA - IRIS)
- William Bisson (Inotiv, contractor supporting RoC)
- Dave Allen (ICCS)

National Institute of Environmental Health Sciences
Division of Translational Toxicology

Annexes

ToxPi



Marvel SW, To K, Grimm FA, Wright FA, Rusyn I, Reif DM. ToxPi Graphical User Interface 2.0: Dynamic exploration, visualization, and sharing of integrated data models. BMC Bioinformatics. 2018 Mar 5;19(1):80.

Reif DM, Martin MT, Tan SW, Houck KA, Judson RS, Richard AM, Knudsen TB, Dix DJ, Kavlock RF. Endocrine profiling and prioritization of environmental chemicals using ToxCast data. Environmental Health Perspectives. 2010. 118(12):1714-20.

Table 1. Summary of the Data Sets Used[a]

data set	compounds	assays	filled
adrenergic	1731	5	37.5%
kinase	13998	159	6.3%

[a]The table shows the data sets, the number of compounds and assays each contains, and the proportion of the compound–assay values that are filled.

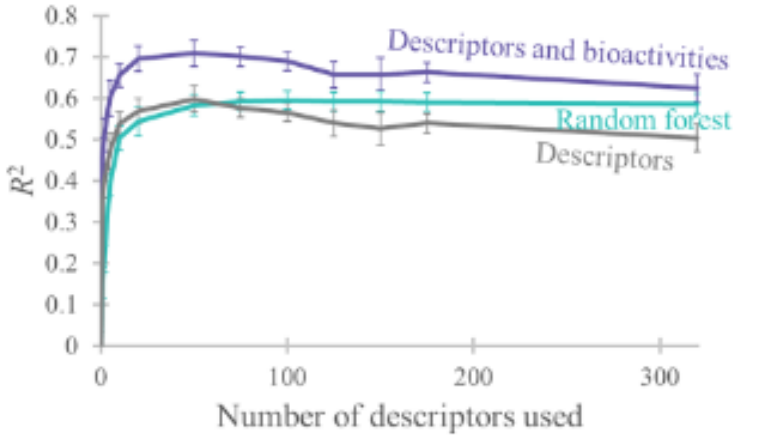


Figure 3. Coefficient of determination for predicting the activity of the adrenergic receptors with number of chemical descriptors. The magenta line is when the neural network is trained with both the activities and descriptors present, the gray line with just the descriptors, and the cyan line is for random forest. Error bars represent the standard error in the mean R^2 over 5-fold cross-validation.

Bioactivities increase performances

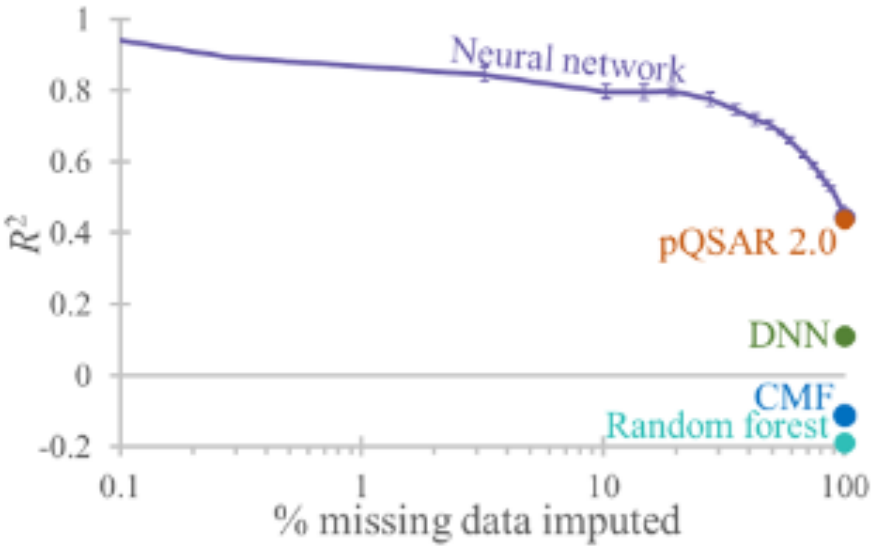


Figure 4. Coefficient of determination for predicting the activity of the clustered Kinase data set with percentage of data predicted. The cyan point is for the random forest approach, the blue point is the collective matrix factorization (CMF) method, the dark green point is the deep neural network (DNN) approach, the orange point is the profile-QSAR 2.0 method, and the purple line is the neural network proposed in this work. The purple line shows that the accuracy of the neural network predictions increases when focusing on the most confident predictions, at the expense of imputing only a proportion of the missing data. This confirms that the reported confidences in the predictions correlate strongly with their accuracy. Error bars represent the standard error in the mean R^2 value over all 159 assays and where not visible are smaller than the size of the points.

Whitehead, T. M., Irwin, B. W. J., Hunt, P., Segall, M. D., & Conduit, G. J. (2019). **Imputation of Assay Bioactivity Data Using Deep Learning**. *Journal of Chemical Information and Modeling*, *59*(3), 1197–1204. https://doi.org/10.1021/acs.jcim.8b00768

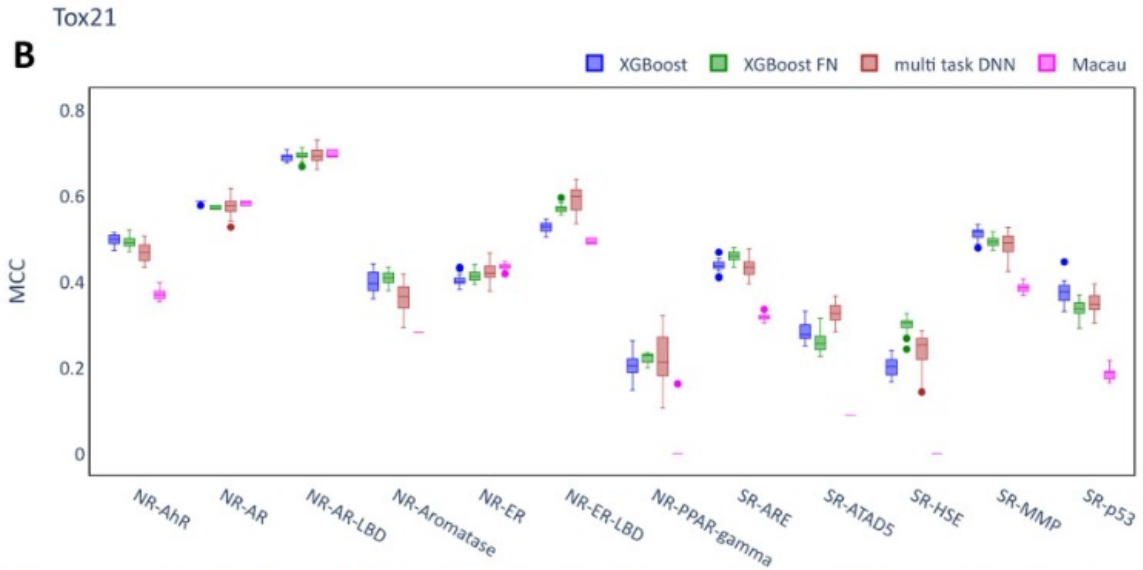


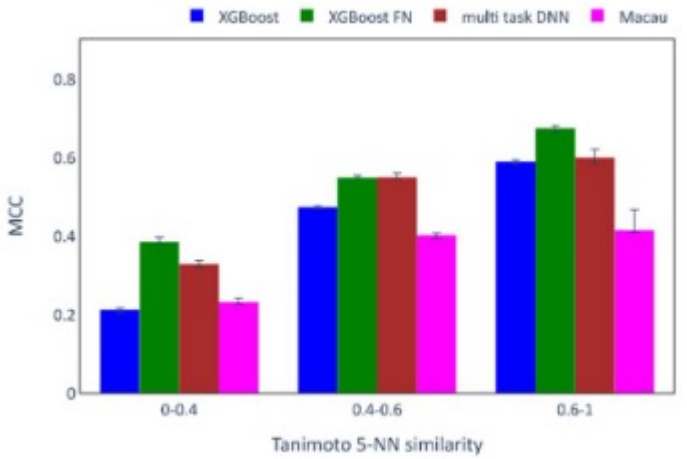
Fig. 3 Performance of multi-task QSAR models. **A** Ames dataset. **B** Tox21 dataset. Each box summarizes the MCC scores of 20 independent runs of the model on the test set with identical hyperparameters but differing random seeds. The XGB models (best performing single task QSAR model) are shown as a baseline model. Only the best performing Feature Net model (XGB-FN) is included.

- Classification model (active no active)
- Test different machine learning
- Single vs multitask (several endpoint predict with one DL model)
- Compound similarity

Tox21

Assay name	Number labels (proportion)	Proportion actives
NR-AhR	6810 (0.84)	0.12
NR-AR	7460 (0.92)	0.03
NR-AR-LBD	6991 (0.86)	0.03
NR-Aromatase	6009 (0.74)	0.05
NR-ER	6367 (0.79)	0.11
NR-ER-LBD	7199 (0.89)	0.04
NR-PPAR-gamma	6752 (0.83)	0.03
SR-ARE	6121 (0.76)	0.16
SR-ATAD5	7326 (0.91)	0.04
SR-HSE	6794 (0.84)	0.05
SR-MMP	6074 (0.75)	0.15
SR-p53	7049 (0.87)	0.06
overall	8090 (0.83)	0.07

assay as well the proportion of active labels for the given assay. The last row



In conclusion, multi-task imputation models have the potential to improve the performance of QSAR models used in practice and to extend their domain of applicability to make predictions for dissimilar molecules.

Walter, M., Allen, L. N., de la Vega de León, A., Webb, S. J., & Gillet, V. J. (2022). Analysis of the benefits of imputation models over traditional QSAR models for toxicity prediction. *Journal of Cheminformatics*, *14*(1). https://doi.org/10.1186/s13321-022-00611-w